Academic Research in East Asia
ISSN: 2519-3392
Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

基于词频统计的林黛玉对话语言特色研究

李坤梅[1]

【摘要】在《红楼梦》中，作为小说女主之一的林黛玉因其个性化的语言特色而备受学者关注和喜爱。目前，对她的语言特色的研究主要是基于小说文本的案例分析，少有研究者采用将质性分析与量化研究相结合的方式，且由于通用分词工具的局限性，白话文本的量化分析在实际应用中，尤其是在分词和情感计算方面，目前还面临着巨大挑战，因此鲜少有人用基于词频统计的方式对白话文本的语言特色进行客观、系统、量化分析。本文将基于在线文本分析软件微词云，通过自建林黛玉话语语料库，对《红楼梦》中林黛玉的对话话语进行词频统计和分析，为研究其语言艺术特点提供量化佐证，同时也为古典文学研究提供一个新的思路。

【关键词】林黛玉；语言特色；词频；微词云；红楼梦

1 李坤梅，广州南方学院，研究方向：语言研究，话语分析，教学研究。邮箱：likm1@nfu.edu.cn.
本文是教育部产学合作协同育人项目“新文科建设背景下综合英语课程与跨学科人才模式融合研究”（项目编号：2508052020）和“数字化转型驱动下英语写作课程思辨能力培养的混合式教学创新（项目编号：2509300734）”阶段性成果。

Academic Research in East Asia
ISSN: 2519-3392
Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

A Study on Linguistic Features of Lin Daiyu's Discourse Based on Word Frequency

Li Kunmei[1]

【**Abstract**】In the classic novel *Dream of the Red Chamber* (*Hong Lou Meng* in Chinese), Lin Daiyu, one of the novel's heroines, has captivated scholarly attention with her distinctive linguistic features. Existing researches on her linguistic features are predominantly conducted through textual case analysis. Few scholars have adopted an integrated approach combining qualitative analysis and quantitative study. What's more, due to the limitations of general-purpose word segmentation tools, quantitative analysis of vernacular texts (Bai Hua Wen in Chinese) are facing significant challenges in practical quantitative research applications, for example in word segmentation and sentiment analysis, thus resulting in difficulties in objective, systematic and quantitative researches into linguistic features of vernacular literature. This study, based on a self-built corpus of Lin Daiyu's dialogue, will make use of the online text-analysis tool MiniWordCloud to conduct analysis of linguistic features in Lin Daiyu's dialogues. This study aims to provide quantitative evidence for exploring her linguistic artistry while offering a novel methodology for classical literary research.

【**Keywords**】Lin Daiyu; Linguistic Features; Word Frequency; MiniWordCloud; Dream of the Red Chamber

1 Li Kunmei, Guangzhou Nanfang College. Research interests: linguistic studies, discourse analysis, pedagogical research. Email: likm1@nfu.edu.cn. This work is supported by the 2025 Collaborative Education Project of Industry-Academia Cooperation (grant number: 2508052020): "Research on the Integration of Comprehensive English Curriculum and Interdisciplinary Talent Cultivation Models under the New Liberal Arts Initiative" and "Fostering Critical Thinking in English Writing Through Blended Learning: A Digital Transformation Perspective (grant number: 2509300734)".

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

引言

《红楼梦》是中国四大名著之一，是中国古典文学巅峰之作。一般认为其前 80 回由清代作家曹雪芹所创，后 40 回由同时期人高鹗续写。《红楼梦》有“中国封建社会百科全书”之称，具有极高的艺术成就和文化价值。小说在人物塑造和语言艺术方面有着独特的魅力。小说里的林黛玉因为鲜明的人物个性和语言特色成为很多人心中《红楼梦》里最出彩的女性人物之一。《红楼梦》本主要有 120 回的“程本”和 80 回的“脂本”。本研究所采集的林黛玉对话语料来自上海大学 2015 年 7 月第一次出版 2021 年 1 月第 10 次印刷的《红楼梦》，共 120 回，共计 90 万余字。

在《红楼梦》中，林黛玉的对话语料共计 16500 余字，在小说中所占的比例并不高。她的话语出现在第 3 回到第 98 回，长达 24 回里她并未实际出场。她在小说第三回初入贾府便以出色的语言艺术“惊艳众人”。她的第一句话出现在书中第三回，原文是：“我自来是如此，从会吃饮食时便吃药，到今日未断，请了多少名医修方配药，皆不见效。那一年我三岁时，听得说来了一个癞头和尚，说要化我去出家，我父母固是不从。他又说：‘既舍不得他，只怕他的病一生也不能好的了。若要好时，除非从此以后总不许见哭声；除父母之外，凡有外姓亲友之人，一概不见，方可平安了此一世。’疯疯癫癫，说了这些不经之谈，也没人理他。如今还是吃人参养荣丸”（曹雪芹，高鹗，2021:20）。初次登场，年幼的黛玉在众多长辈们面前丝毫不露怯，言谈举止间将大家闺秀的大方得体表现得淋漓尽致。她的最后一句话出现在书中第九十八回，彼时宝玉正在娶亲，而她则已油尽灯枯进入弥留之际，留下临终未尽之语：“宝玉，宝玉，你好……”，有人说这句未尽之语是本书最具张力的留白艺术，它像一面镜子，照见的是解读者的心相：怨者见其恨，悟者见其慈，痴者见其痛（刘再复，2009）。

林黛玉的语言是《红楼梦》中最具辨识度的语言艺术，其对话本身便是一个值得深入探究的复杂文本。它呈现出一种“带刺玫瑰”式的独特张力：尖刻与柔情并存，既是一位诗意灵魂在礼教秩序下内心挣扎的微观实录，也为探析个体如何凭借言语在结构性压抑中寻求自由，提供了一个经典的文学样本。

一. 文献回顾

《红楼梦》是“中国封建社会百科全书”，是一部伟大的语言巨著。自诞生以来，其独具匠心的语言表达一直受到人们的广泛关注。就以量化方法研究《红楼梦》而言，其实

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

上个世纪 80 年代就已经有学者开始尝试了，比如《从数理语言学看后四十回的作者》（陈大康，1987），《〈红楼梦〉成书新说》（李贤平，1987），但这些研究主要是以计算机作为辅助对《红楼梦》的作者进行识别和判断。进入 21 世纪以来，随着计算机时代的到来，人们对《红楼梦》的研究从传统文学的深耕，逐渐转到跨学科视角的多元创新，涌现了一些学者结合计量语言学、统计学以及数字人文方法，对《红楼梦》从作者考证、人物分析、文本风格、情感计算等多方面进行研究。施建军（2011）在《基于支持向量机技术的〈红楼梦〉作者研究》中，以计算机人工智能技术为辅助，以小说中的文言虚字频率为特征向量，对《红楼梦》进行分类研究，从技术上证明《红楼梦》前 80 回和后 40 回存在写作风格差异。胡翠婷（2019）的《基于词频计量统计的林黛玉性格分析》通过对林黛玉 23 首诗词进行诗词词频和动词词频统计，并通过情感极性标注验证林黛玉的情感分布情况。刘颖、肖天久（2014）的《〈红楼梦〉计量风格学研究》通过研究词长分布、统计独有词和对高频词进行聚类分析等方式，发现了小说前 80 回和后 40 回存在句式结构差异。郑佳莉等（2022）在《基于词频分析的 K-Means 特征聚类算法的〈红楼梦〉作者分析》一文中，提出以基于词频分析的 K-Means 特征聚类算法来分析存疑文献的方法，通过聚类算法分析后，作者认为前 80 回和后 40 回是不同作者所创。总之，随着时代的发展，如何更加科学地进行语言研究一直是现代学者们的努力方向和目标，学者们采用定量分析的方法为语言进一步发展和研究提供了思路（江铭彪，2023）。

以小说中林黛玉的语言来说，学界普遍认为她的语言兼有风趣幽默、端庄典雅、尖刻而又不失率真的特点（张宜平，2017；龙锦婷，2021），但这些研究都主要是基于小说文本的案例分析，鲜少有人采用定量研究的方法从数据驱动的客观视角出发对林黛玉的语言特色进行客观、系统、量化分析。而且目前白话文本的量化分析在实际应用中，尤其是在分词和情感计算方面，还面临着巨大挑战。本研究将基于在线文本分析软件微词云，通过自建林黛玉话语语料库，对《红楼梦》中林黛玉的对话话语进行词频统计和分析，为研究其语言艺术特点提供量化佐证。基于此，本研究提出以下研究问题：

1. 从微词云在线系统的分词统计结果来看，林黛玉对话话语有什么语言特色？

2. 将系统分词的统计结果和人工校对后的统计结果进行对比研究后，林黛玉对话话语语言特色有无明显语言特色变化？

3. 本研究对尝试进行质性分析和量化分析相结合的古典文学研究有何启示意义？

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

二. 数据来源与研究方法

2.1 数据来源

本研究所采用的语料来上海大学 2015 年 7 月第一次出版 2021 年 1 月第 10 次印刷的《红楼梦》，共 120 回，共计 90 万余字。为了系统完整地呈现林黛玉的语言特色，为确保语料的完整性和准确性，也考虑到古白话对统计准确性的影响，本研究里所有林黛玉的对话语料全是通过人工提取整理而来，没有用 python 以及其他计算机辅助手段进行语料整理，通过人工筛选和整理校对后，总共提取林黛玉纯对话语料 16110 余字，对话条数 553 条。特别说明的是，统计的时候，林黛玉的联句算入在统计语料内，因为联句出现在和大观园姐妹们的互动场景中，但她写的诗词不算在对话语料内，她的诗作如《葬花吟》、《秋窗风雨夕》、《问菊》等是她才华的体现，但是并不出现于对话的场景，因此本研究在研究对话语言特色的时候，没有把她的诗作纳入在内。

2.2 研究方法

微词云（MiniWordCloud）是一款在线文字云和词云生成工具，具有强大的文本处理能力、个性化的设计功能、专业的文本分析功能和多样化的导入导出格式。对于文本分析来说，它的核心优势之一是实现零代码操作，只需要直接粘贴或者上传 txt 格式或者 docx 格式的文本，然后进行简单的设置即可开始自动分词（内置 Jieba 分词，也可以导入自定义词典），快速进行词频统计，生成可视化文本关键词等。近年来，因为其便捷简单的在线功能、强大的文本处理能力和优秀的输出效果，它已逐渐成为国内一些学者在做文本分析、报告制作、数据可视化等工作时的得力助手。

为了破解《红楼梦》中林黛玉对话话语中流动的基因密码，本研究融合统计计算和人文学养，利用微词云对林黛玉对话语料进行初步中文分词和词频统计，但考虑到目前白话文在量化分析时正面临着分词准确性、情感复杂性和语境依赖性的巨大挑战，本研究所有的语料都事先经过人工预处理，尤其是在分词的时候，专门建立了匹配的自定义词典和同义词词典，并且在微词云实现分词后，仔细核对打标词表和特征词表，逐条进行人工修正和匹配，以最大可能确保数据统计的准确性和研究的可信度。

三. 词频统计与对话语言特色

3.1 基础信息及统计数据解读

角色的语言是塑造其形象、推动情节、表达情感的核心载体。研究林黛玉的语言特色，

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

我们首先将预处理好的对话文本转换成 txt 格式。同时，针对该文本，要整理好一份尽量详细的自定义词典，以破解未录入词困局，确保分词的准确性。比如以小说《红楼梦》为例，“芦雪庵”“侍书”“蕉下客”“葬花吟”“携蝗大嚼图”等众多复合词需要添加进自定义词库成为专有词库。自定义同义词词典的主要作用是消解语义多样性陷阱，比如在林黛玉对话语料里，“死了”跟“去世”“去了”“死的”等具有语义关联，“林姑娘”跟“林黛玉”“黛玉”“颦儿”“潇湘妃子”“林妹妹”等具有人物关联。

接下来，打开微词云在线系统，分别导入 txt 格式的林黛玉对话语料、自定义词典、同义词词典后，选择“去重”和单词长度 >=2，点击“下一步”，系统就开始自动分词和计算，快速生成基础信息（见表 1）、单词分布情况、词云图、特征词表、词性柱状图、网络关系图等。

表 1　林黛玉对话话语的基础信息表

指标	数值	说明
总字数	16,566	文本总字数
总词数	2,749	分词后总词数
特征词数	1,356	具有统计意义的特征词数量
平均句长	4.97 词 / 句	每句平均词数（总词数 / 总句数）
词密度	67.88%	特征词数占总词数的比例
有效条数	553 条	有效分析条目数量
可下载资料	打标词表、特征词表等	配套分析资源

表 1 是系统生成的通用基础信息表，直观呈现语料基本信息。总对话字数超过 16000，从字数上看角色在小说中对话量还是可观的，从发言量可看出角色在小说中的地位和重要性。整本小说中林黛玉对话语料中的总词数为 2749 个（单词长度 >=2，含重复词、停用词的情况下），特征词数为 1356，是剔除掉重复词和停用词后（系统内置停用词表）的有效词数量，特征词占比为 1356/2749，大约 49.3%，说明该文本信息浓度比较高，冗余词比较少。特征词是林黛玉语言风格、思想感情和身份背景等信息的独特标志，49.3% 的占比说明文本中有一半是具有林黛玉个人特色的词汇，语言高度个性化。这些特征词是我们了解林黛玉个性化语言风格、情感表达主题以及身份背景的语言指纹。

平均句长为 4.97 个单词，每句话平均包含的词语数为将近 5 个词 / 句，低于中文的平均对话话语句长，在书面语中属于较短的句子，这意味着尽管文本是文言白话的混合体，

ISSN: 2519-3392　　DOI: 10.64699/25/ADOL7763

林黛玉的对话更接近日常对话的节奏，口语化、非正式或非规范表达明显，情感表达倾向直接或急促。平均句长偏短还说明整个对话语料比较碎片化，因为里面有许多短句、口语化表达。结合表 1，整个对话语料共计 553 条，但是依据标注词表进行统计，字数 =<10 条数的居然达到 175 条，占比为 175/553，大约 31.6%，在这些短句中，词数分布为 0 个 / 句 --3 个 / 句不等，且多数为 1 个 / 句。总之，从较短的句长这一特点来看，我们可以发现主角情感表达浓烈，喜欢直接、感性的表达，符合她敏感多思、情绪易波动的性格。

词密度（lexical density）通常指承载主要意义的实词占比。67.88% 的高词密度，说明 67.88% 的词为有效词。这些有效词是充满实质性的情感词汇和意象。词密度较高说明林黛玉的话语信息密度大，情感表达集中，语言表现能力很强。

微词云非常方便数据的输出和下载，比如表 1 最末一列，就清晰的显示了可以下载的文档或资料，方便研究者进行信息比对和校正。除了表 1 所列的数据下载信息外，其它大部分核心数据都可以免费下载，比如特征词表、词云图、单词分布图、词性柱状图、网络关系图等。

总之，据初步数据显示（表 1），林黛玉的对话语言有用词独特、情感表达细腻的特点，反映其敏感多思的性格特征。语言是灵魂最直接的外化，这些数据从量化角度为我们解读林黛玉这一经典文学人物提供了独特的视角。

3.2 词频统计及数据解读

图 2 是在微词云在线分词后形成的中文通用分析词频统计图，将整个林黛玉对话语料库中有效词的分布情况实现量化，按比例从高往低依次为：名词 29.03%，动词 24.55%，代词 17.02%，副词 8.55%，数词 6.55%，形容词 5.75%，习用语 2.87%，人名 2.22%，成语 1.71%，其它 1.8%，总词数 2749 个，共计 1356 个特征词。

表 2　基于微词云的林黛玉对话语料词频统计表（人工校正前）

词类	占比	词频	大致占比图
名词	29.03%	798	█████████████████████████ ██████ 29.03%
动词	24.55%	675	█████████████████████████ 24.55%
代词	17.02%	468	███████████████ 17.02%
副词	8.55%	235	████████ 8.55%
数词	6.55%	180	██████ 6.55%

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

（接上表）

形容词	5.75%	158	■■■■■ 5.75%
习用语	2.87%	79	■■ 2.87%
人名	2.22%	61	■■ 2.22%
成语	1.71%	47	■ 1.71%
其它	1.80%	48	■ 1.80%
总计	**100%**	**2749**	

但是在进一步核对系统分词后输出的特征词表后，我们发现系统在分词和统计的时候出现了一些明显的错误。依据表 2 词频统计表，我们按照排名先后顺序依次逐条对词频表进行人工识别，将出现分词错误或者分类错误的特征词进行汇总。整体来说，代词和数词错误较少，这应该与现代汉语和白话文在代词和数词方面的差异不大有关，代词错误率相对较高，约 3.4%，数词错误率仅 0.5%。其它（地名专名）项、人名、成语、形容词等项错误率明显高，这些错误凸显出目前白话文分词的难点和困局。汇总详情见表 3：

表 3　系统分词错误统计表（人工校正前）

错误类型	错误个数	总频次	错误率	错误程度
名词	96	128	16.0%	高
动词	40	56	8.3%	中
代词	11	16	3.4%	低
副词	11	12	5.1%	低
数词	1	1	0.5%	极低
形容词	27	43	27.2%	高
习用语	3	3	3.8%	低
人名	16	18	29.6%	高
成语	7	9	19.1%	中
其他	34	44	91.7%	！极高

根据上面表 3 数据，各类语言错误呈现显著差异：其他类别错误率异常突出（91.7%），需深入排查原因。人名（29.6%）和形容词（27.2%）分词错误率偏高，需重点关注；名词错误率达 16.0%，也是核心问题领域。动词（8.3%）和成语（19.1%）错误率中等，需适

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

度关注；代词（3.4%）、副词（5.1%）及习用语（3.8%）错误率较低。数词错误率极低（0.5%），几乎可以忽略。整体而言，其他类别、人名、形容词和名词错误是主要关切点。

在上面表 3 中，共计有 246 个分词错误或分类错误的特征词，合计 330 频次。从系统导出打标词表分析和统计后发现，绝大多数的统计错误出现在特征词的分类错误方面，占比约 90%。分词错误的特征词共计 30 个，合计 34 频次，分词错误比例为 34/330，约为 10.2%。明显的分词错误，比如“得宝”“干！”，这样的错误很容易识别和校正，但是“隐性的”分词错误往往容易被忽略掉，因此在研究白话文本的时候，为了保证数据的准确和可靠，研究人员需要花费大量的时间定位到原文文本进行核验。在 30 个错误分词里，有 17 个是属于比较明显的错误分词，剩下 13 个为“隐性”错误分词，即从语义和结构上粗略一看没有明显不妥或突兀之处，但是定位到词语出现的具体文本后，会发现明显错误分词，比如“怪道人”一词，出自《红楼梦》第 59 回林黛玉对丫鬟的赞美之语，原文是“怪道人赞你的手巧，这顽意儿却也别致”（曹雪芹，高鹗，2021: 469）。此处的“怪道人”应该分成“怪道 + 人”才合理，“怪道”作副词用，表示难怪的意思。

3.2.1 基于词类的典型统计错误分析

基于词类的典型错误分析不仅是探究系统分词错误成因的关键，更是确保校对工作有效性的基石。唯有通过精准识别并校正错误，后续的文本和数据分析质量才能有坚实的基础保障。

从表 3 来看，名词类的不当分词数量最多，约占比 38.4%。错误分词 13 个，其中明显错误分词 7 个，“隐性”错误分词 6 个。“显性”错误分词在语义上往往比较模糊，或者表达结构上有缺陷，以错误分词“上人”、“母弱弟”为例，原文中，“上人”出自书中第 67 回，原话是“你要你只管说，不必拉扯上人”（曹雪芹，高鹗，2021: 543）。而“母弱弟”，则是出于第 35 回，原文是“然你虽命薄，尚有孀母弱弟”（曹雪芹，高鹗，2021: 262-263）。相对而言，“隐性”错误分词比较难识别，最容易被判定为分类错误，必须要定位到具体的小说文本且结合语义才能正确判断。

动词组出现统计错误的词数为 40 个，总频次为 56，占比约 16.8%，隐性错误分词仅 4 个。以典型的隐性错误分词“年少”和“出新”为例，这两个词在现代汉语里是常用词，但是在原文里，应该是“比 + 旧年 + 少”和“兴出 + 新文”这样的划分结构。被划分为动词的显性错误分词“比不”情况比较特殊，是一个多频次出现的特征词，在原文中分别出自第 22 回“你不比不笑，比人比了笑了的还利害呢！”（曹雪芹，高鹗，2021: 162）和第 28 回“比不得宝姑娘，什么金什么玉的，我们不过是草木之人！”（曹雪芹，高鹗，2021:

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

217）和第 82 回“二爷如今念书了，比不得头里”（曹雪芹，高鹗，2021: 674）。第一处“不比不笑”应该分词为“不比 + 不笑”，第二处和第三处应该是“比不得 +XX”结构，因此虽然可以很容易判断“比不”是错误分词，但是校正的时候还需要根据具体情况具体处理才行。

副词组里，共有 11 个分词或分类错误，属于错误相对较少的类别了。“不齐”和“原为”是典型的白话文。“不齐”不是现代文里的“不整齐”的意思，而是“说不齐”，“原为”的结构是“原 + 为”，意思是“原来是因为”。“更次”如果不结合文本，容易望文生义为“更替的次序”、“更加次之”等意思，通过文本定位找到第 52 回原文，“昨儿夜里好了，只嗽了两遍，却只睡了四更一个更次，就再不能睡了”（曹雪芹，高鹗，2021: 406），才发现此处“更次”在分词上并无不妥，只是此处“更次”当归属名词类，表示时间概念。

表 3 的形容词组一共有 27 个 43 频次的错误统计，其中 5 个错误分词。“干！”是明显错误分词，没有满足单词长度 >=2 的设定要求。在微词云词频统计表中，形容词频次最高的是消极情感词汇“死了”，一共出现 9 次，但有趣的是，“死了”在文本中的结构基本都是“某人 + 死了”，由于“死”是动词，描述的是某人从生到死的状态变化或结果，因此“死了”当属动词范畴。

代词、习用语、数词、人名和成语出现的分词或分类错误从整体来说比较少，总共 38 个，由表 3 可见，分别是代词 11 个，习用语 3 个，数词 1 个，人名 16 个，成语 7 个。习语部分，以“无风”为例，原文是“无风仍脉脉”（曹雪芹，高鹗，2021: 387）。此处“无风”只是表达“没有风”这个普通的意思，因此不属于习语。“人不知”这个分词出自原文“只是今人不知，误作俗字用了”（曹雪芹，高鹗，2021: 623-624），由此可见此处结构应是“今人 + 不知”。

由于白话文和现代汉语在数词的使用上并没有明显差别，因此系统在数词分词方面错误很少，只有一处：“五儿”，在小说中是丫鬟的名词，因此当归属人名。人名组中，“会子”出自“他两个再到不了一处，若到一处，生出多少故事来。这会子一定算计那块鹿肉去了。”此处“这会子”相当于“这会儿”。另外“威福”出自原文“差不多的人就早作起威福来了”（曹雪芹，高鹗，2021: 498）。在此句中，“威福”含明显贬义，虽然在现代汉语中不常用了，但是“威福”作为文言词语出现在《红楼梦》中是合理的，只是不属于人名。在其它类中，统计出来的分词或分类错误数量整体来说是比较多的，仅次于名词，达 34 个之多，共计 44 频次，占比约 13.2%。

总之，在基于系统分词结果对分词、分类错误依照词类进行的典型案例分析后，我们

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

发现在对《红楼梦》中的林黛玉对话语料进行分词与词类标注分析中，名词类错误占比最高（38.4%），其中“显性”错误（如“上人”“母弱弟”）多因语义模糊或结构缺陷；“隐性”错误（如“年少”“出新”）需结合上下文才能识别。动词组错误占比 16.8%，以“比不”等结构误判为代表，需根据语境校正。副词组错误较少。形容词组中“死了”虽频次高，但词性尚待商榷。代词、习用语等类别错误较少，但“威福”等词要考虑其在文言语境下的合理性。其他类错误占比 13.2%，以“步田地”等结构误分为主。整体而言，白话文系统分词、分类错误原因多样化，需结合具体文本与语义，且隐性错误识别难度较大。

3.2.2 人工校正后的统计对比

在表 3 的基础上，我们依据系统生成的打标词表，通过人工逐条校对、修正与完善，对出现系统分词错误及统计错误的特征词进行了精准识别与分类。经此过程，我们生成了人工校正后的新特征词统计表（表 4）。该表以量化数据为基础，为接下来深入分析林黛玉的语言特色提供了更为可靠的数据支撑。

表 4　人工校正前 VS. 人工校正后的系统分词错误统计对比表

词性	人工校对前（个数 / 频次）	人工校对后（个数 / 频次）	变化趋势（增减）	具体情况（个数 / 比例）
名词	96 /128	89/99	↓	-7/7.3%
动词	40 /56	81/121	↑	+41/102.5%
代词	11/16	1/1	↓	-10/90.9%
副词	11/12	28 /46	↑	+17/154.5%
数词	1/1	7/9	↑	+6/600%
形容词	27 /43	30/40	↑	+3/11.1%
习用语	3/3	18 /21	↑	+15/500%
人名	16 /18	8/10	↓	-8/50%
成语	7/9	12/12	↑	+5/71.4%
其它	34 /44	2/3	↓	-32/94.1%
总计	**235/320**	**276/340**	**↑**	**+41/17.4%**

对出现系统分词错误的词语进行人工校对后，通过表 4 的数据对比，不难发现大多数词类都呈现显著变化趋势：名词（-7/7.3%）和人名（-8/50%）的个数和频次虽均有所下降，但是幅度并不算大；动词（+41/102.5%）、副词（+17/154.5%）、数词（+6/600%）、习用语（+15/500%）及成语（+5/71.4%）的个数和频次均有大幅上升，其中数词和习用语的

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

增幅尤为突出；形容词（+3/11.1%）个数和频次有小幅增长，而代词（-10/90.9%）和其它类词（-32/94.1%）的个数和频次则大幅减少。考虑到基数问题，名词（-7/7.3%）和形容词（+3/11.1%）属于实词中波动不大的部分，而动词（+41/102.5%）和副词（+17/154.5%）则波动明显，由此可见，在词频研究中，研究者对白话文动词和副词方面要多加关注。整体来看，人工校对后总词数和频次都有明显增加。

需要特别说明的是，在人工校对的过程中，会出现一些汉语构成相似，但是基于文本进行分词后词类归属不同的情况。比如“立足”和“立意”，都是“立 +X”结构，虽然二者在现代汉语中都可做动词和名词，但在分词的时候，回归到具体的文本中后，二者的词性出现差异。“立足”在文中有两次出现，都是来自第 22 回，原文是“你那偈末云，‘无可云证，是立足境'，固然好了，只是据我看，还未尽善。我再续两句在后。”和“无立足境，是方干净”（曹雪芹，高鹗，2021：164）。在上面的话语中，“立足”表示“站立”的意思，虽然动词属性强，但此处作定语修饰“境”，“无立足境”可拆分为“无 + 立足 + 境”，意指没有立足的地方，所以人工校对后把“立足”纳入形容词类。“立意”出自文中第 48 回，原文是“正是这个道理，词句究竟还是末事，第一立意要紧。若意趣真了，连词句不用修饰，自是好的，这叫做‘不以词害意’”（曹雪芹，高鹗，2021：372）。在此句中，“立意”无疑是名词性的了。

整合表 2 和表 4，我们形成了人工校对前和人工校对后的总特征词统计表（表 5），我们发现经过人工校对和完善后，增减明显的主要是名词、动词、副词和其它地名专名项。名词和其它项特征词下降趋势明显，分别下降 1.3% 和 1.5%，而动词和副词特征词则增加比较明显，分别增长 2.1% 和 1.15%。而作为实词的形容词则变化并不明显，只略有下降。具体对比请参照表 5：

表 5　人工校对前 VS. 人工校对后的林黛玉对话语料词频统计对比表

词性	人工校对前	人工校对后	变化趋势	增减值
	占比	数量	占比	数量
名词	29.03%（798 个）	27.7%（769 个）	↓	-1.3%
动词	24.55%（675 个）	26.6%（740 个）	↑	+2.1%
代词	17.0%（468 个）	16.3%（453 个）	↓	-0.7%
副词	8.55%（235 个）	9.7%（269 个）	↑	+1.15%
数词	6.55%（180 个）	6.8%（188 个）	↑	+0.25%
形容词	5.75%（158 个）	5.6%（155 个）	↓	-0.15%

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

（接上表）

习用语	2.87%（79 个）	3.5%（97 个）	↑	+0.63%
人名	2.22%（61 个）	1.9%（53 个）	↓	-0.32%
成语	1.71%（47 个）	1.8%（50 个）	↑	+0.09%
其它	1.8%（48 个）	0.3%（7 个）	↓	-1.5%
总计	**100%（2749 个）**	**100%（2781 个）**	**+1.16%**	**+32 个**

3.3 基于词类的语言特色解读

表 5 统计数据显示，从整体排名来看，不管是人工校对前还是人工校对后，各类词的排名整体并无变化。基于表 5 的最新统计数据，我们算出最新的词密度值和平均句长。词密度 = 实词数量 ÷ 总词数 ×100%，实词数量 = 名词数量 + 动词数量 + 形容词数量 + 副词数量，由表 5 可知，实词数量 =769+740+155+269=1933，因此最新词密度 =1933 ÷ 2781 ≈ 69.51%，比人工校对前的词密度值 67.88% 增加了 1.63%。平均句长 = 总词数 ÷ 总句数，因此平均句长 =2781 ÷ 553 ≈ 5.029，比人工校正前增加了 0.059 词 / 句。

因此，基于我们最新的统计数据，词密度 69.51%，平均句长 5.029 词 / 句，与人工校正前系统分词结果没有太大出入，印证了我们对林黛玉语言风格的整体判断：语言凝练，信息密度高，说明表达者聪明多才；句子短促，短句占比高，说明表达者情感表达直接敏感，易多愁善感。

从词类占比来说，名词和动词合计占比达 54.26%，说明整个对话语言由名词和动词主导。代词使用高频，占比达 16.29%，显示说话者自我意识很强，对话多聚焦具体事物、动作及人物关系。整体来说，对话中文化元素丰富，含有“文化负载词”的习用语、成语以及部分人名专名的使用合计占比达到 7.2%，高于形容词的比例，显示说话者不光自我意识明显，还富有诗书才华。

据表 5 显示，名词占比最高，达 27.7%，说明对话语料中，说话者侧重对具体对象、事物或概念的表达。对话中生活细节丰富，比如“葬花”，“咏菊”等情节，体现说话者对自然和死亡的敏感和关注。另外，从总特征词表排名来看，有很多排在前面指向人的名词，如“老太太”“哥哥”“姐姐”“姑娘”等，这些名词或称呼的密集出现侧面反映出大观园中群像丰富程度。“姨妈”“舅母”等长辈亲属称谓的高频出现则反映出林黛玉的社交礼仪与情感张力，凸显封建家族的复杂姻亲关系。

动词占比为 26.6%，只略低于名词，体现出对话中话语输出者动态感和行为指向性比

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

较强，擅长决策或推动事情往前发展，说话者是比较干练聪敏的行事风格。高频动词里有许多认知动词和否定词，如“没有”“知道”“不知”“不能”“不用”“不好”等，让人仿佛看到了人物内心的矛盾以及封建礼教压抑下人物在命运面前的无力感，整体让人感觉到消极意味比较浓，符合人物敏感多疑、情感波动较大的性格。

代词占比高达 16.3%，甚至比形容词和副词占比之和还高。从词频表看，人称代词里前五排名：“你们”>“咱们”>“他们”>“我们”>“自己”，其中“你们”的词频远远高于其他词，显示对话指向性明确，但是“咱们”次高频，又显示说话者的包容性与亲和力，涉及自我指涉的“自己”相对低频，凸显反身性弱化。指示代词词频排名：“那里”>“这里”>“这个”>“这些”>“这样”，“那里”>“这里”，显示说话者空间指向偏向“远指”。疑问代词词频排名：“什么”>“怎么”>“为什么”>“怎么样”>“如何”，其中“什么”以 76 的绝对高频碾压排名第二词频为 33 的指示代词“那里”，同时从词频上看，“什么”>“怎么”体现说话者以信息询问为主的语言特色。这些极高词频的代词背后综合体现的是林黛玉人际指向和空间指向的信息，超高词频的疑问代词更暴露其“多疑”的特质。

副词占比约 10%，是一个比较平衡的比例，说明修饰动作或情感强度的词语比例适中，结合微词云导出的副词特征列表，发现有许多程度副词和修饰性副词值得关注，如“越发”“真真”“到底”“原来”“就是”“自然”“已经”等，从一定程度上证明了林黛玉说话直接率真的特点，容易情感外露，但同时丰富的修饰性副词也让她的话充满了细节感。数词占比偏高，说明说话者描述具体准确，反应其心思细腻、观察入微的个性。以位居前五高频数词为例，“一句”“一 / 两 / 几个”“一首”“一年”“几处”，这些高频数词反应了说话者在涉及数量、时间、位置等信息表达上的特点和倾向。数词数量的丰富体现了说话者冷静理性的一面，凸显其对精确性与秩序感的追求。

形容词占比偏低，仅 5.6%，说明说话者对描述性语言保持了一定的克制，削弱了主观情感渲染，说话者属于务实风格，整个说话以冷静陈述为主导。另外，形容词少暗示说话者话语节奏比较快，在对话中喜欢用简单直白的词汇，生活化口语化特征很明显。但是特征词表也显示：词语的情感评价出现明显两级分化倾向，即积极词与消极词对比强烈，比如“新鲜”“热闹”“高兴”“极好”“有趣”这些词都带有明显的积极意思，而“不干净”“不巧”“不大”“恍惚”等消极意味明显，反应出说话者语言直接、情感鲜明，可能比较固执或带有抱怨情绪。总之，从微词云的形容词特征词表来看，林黛玉话语中的直率、生活化、情感外露的语言风格鲜明。

习用语虽然占比只有 3.5%，但对话中许多明显口语化的表达，如俗语、口头禅等，既

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

是《红楼梦》语言艺术的精华，也是人物塑造不可缺少的手段。典型的习用语如“抱不平儿”“耳旁风”“嚼舌子”“莫不是”“多早晚”“顽意儿”“少不得”“不中用”“正经人”“心拙口笨”等，习用语是人物灵魂的体现，也是作者艺术匠心的密码。习用语的绝妙使用既体现闺阁对话的真实性，也显示其机锋暗藏或尖刻的一面。

人名虽然占比占比只有 1.9%，但是历史文化人物如“陶渊明”“荆轲”“李义山”“东方朔”“江淹”等都是曹雪芹精心设计的与林黛玉有价值关联的才高命蹇之人，他们都才华横溢却又命途多舛。名字即命运，林黛玉提及的这些历史人物折射她的人格特征、精神追求和悲剧命运，也是其“书香门第”背景的很好体现。成语同人名一样，虽然占比并不高，但是林黛玉对话中提到的文学性成语却很丰富，如“高山流水”“蟾宫折桂”“起承转合”“物伤其类”“纸上谈兵”“卷舒自若”“青女素娥”“鹤山凤尾”等，体现其对话的文学性和社会性，彰显说话者的诗书才华，亦是其“才女”身份的印证。

其它项主要是地名和专有名词，频率较高的有“扬州”“阿弥陀佛”“罗汉”“凹晶馆”“芦雪庵”等。“扬州”代表林黛玉的回不了的过去，是她的“无根之痛”，“凹晶馆”“芦雪庵”跟“潇湘馆”一样是曹雪芹特意构建的空间符号，代表的是说话者的“现在”，是她在贾府“暂居”的主要活动场所，呼应其作为“他者”“寄人篱下”的处境。林黛玉的“阿弥陀佛”多带反讽或自嘲的意味，也是一种“槛外人”的自我投射，以“孤标傲世”对抗“看破红尘”，无情揭露现实社会的虚伪。

四 . 结语

本研究基于在线文本分析软件微词云，通过自建的林黛玉对话语料库，对《红楼梦》中林黛玉的对话话语进行词频统计和分析，将量化分析与质性分析相结合研究林黛玉的语言特色。本研究发现，尽管自动分词在具体词项的归类上存在一定误差，但从词类分布的宏观格局来看，人工校对并未改变名词、动词、代词占据主导地位的整体排名顺序。同时，词密度（从 67.88% 提升至 69.51%）与平均句长（从 4.97 词 / 句微增至 5.029 词 / 句）在校对前后均保持高位稳定，这一结果从量化角度强化了对林黛玉语言风格的核心判断。基于校正后的词频数据，林黛玉的语言特色得以被量化地勾勒出来：高达 69.51% 的词密度和约 5 个词 / 句的平均句长，共同印证了她语言凝练、信息浓度高的特点，反映出其思维的敏捷与才情；而短促的句式和高达 31.6% 的超短句（≤ 10 字）占比，则直观地体现了其情感表达的直接与敏感。

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

从词类占比来说，整个语言由名词和动词主导。林黛玉侧重对具体对象、事物或概念的表达，对话中动态感和行为指向性比较强，擅长决策或推动事情往前发展，行事风格属干练聪敏类型。代词使用高频，显示她自我意识很强，对话多聚焦具体事物、动作及人物关系。副词占比平衡，有许多程度副词和修饰性副词值得关注，这些词体现了她说话率真容易情感外露但又充满生活的细节感。数词占比偏高，说明林黛玉描述具体准确，反应其心思细腻、观察入微的个性。形容词占比偏低，表明她对描述性语言保持了一定的克制，主观情感渲染较弱，属于务实风格，整个说话以冷静陈述为主导。另外，形容词少暗示林黛玉话语节奏比较快，在对话中喜欢用简单直白的词汇，生活化口语化特征很明显。但是特征词表反应词语的情感评价出现明显的两级分化。统计还显示，林黛玉对话中文化元素丰富，习用语、成语等都是林黛玉“黛玉腔”不可或缺的要素。

本研究是一次古典文学研究的探索之旅。由于通用分词工具的局限性，目前白话文本的量化分析在实际应用中还面临着一些挑战，如在分词的方面，为了确保分词的准确性，需要建立相应的专有词典，再如在对词语进行情感分析的时候，需要对一些特殊的词如“忒”“真真”进行专门标注，否则会出现情感极性误判。以量化方式对白话文本进行客观、系统的分析是对古典文学研究的大胆尝试，对我们在新形势下更好地认识、解读我们的经典作品具有十分特别的意义。

【参考文献】

曹雪芹，高鹗 . (2021). 红楼梦 [*Dream of the Red Chamber*][M]. 上海：上海大学出版社 [Shanghai: Shanghai University Press].

陈大康 . (1978). 从数理语言学看后四十回的作者—与陈炳藻先生商榷 [*On the Authorship of the Last Forty Chapters from the Perspective of Mathematical Linguistics: A Discussion with Mr. Chen Bingzao*] [J].《红楼梦学刊》[*Studies on “A Dream of Red Mansions”*]，(1): 293-318.

胡翠婷 . (2019). 基于词频计量统计的林黛玉性格分析 [*Study on Lin Daiyu’s Personality Analysis Based on Word Frequency Measurement Statistics*] [J].《现代语文》[*Modern Chinese*]，(2)：86-92.

江铭彪 . (2023). 文学作品的统计分析 [*Statistical Analysis of Literary Works*] [M]. 中国文史出版社 [China Culture and History Press]，4-7.

李贤平 . (1987).《红楼梦》成书新说 [*A New Thought on the Compilation of Dream of the Red Chamber*] [J].《复旦学报》（社会科学版）[*Fudan Journal (Social Science Edition)*]，(5)：1-36.

Vol.10 No.2 2025
DOI: 10.64699/25/ADOL7763

刘颖，肖天久 . (2014).《红楼梦》计量风格学研究 [*A Quantitative Stylistic Study of Dream of the Red Chamber*] [J].《红楼梦学刊》[*Studies on "A Dream of Red Mansions"*]，(4): 47-65.

刘再复 . (2009). 红楼梦悟 [*Insights into Dream of the Red Chamber*] [M]. 北京：生活・读书・新知三联书店 [Beijing: SDX Joint Publishing Company].

龙锦婷 . (2021). 林黛玉交际语言特色探析 [*Lin Daiyu's Communicative Language Features*] [J].《齐齐哈尔师范高等专科学校学报》[*Journal of Qiqihar Teachers' College*]，(2): 54-58.

施建军 . (2011). 基于支持向量机技术的《红楼梦》作者研究 [*Dream of the Red Chamber Authorship Study Based on SVM Technology*] [J].《红楼梦学刊》[*Studies on "A Dream of Red Mansions"*]，(5): 35-52.

张宜平 . (2017). 林黛玉语言风格的形成原因分析 [*Analysis of the Formative Factors of Lin Daiyu's Linguistic Style*] [J].《文学教育》[*Literature Education*]，(1): 9-16.

郑佳莉，柯小玲，江晓莹等 . (2022). 基于词频分析的 K-Means 特征聚类算法的《红楼梦》作者分析 [*Analysis of the Author of A Dream of Red Mansions Based on K-Means Feature Clustering Algorithm with Word Frequency*] [J].《数据挖掘》[*Data Mining*]，12(1): 73-79.

（责编：杨维忠）